Towards Using Data Driven Lemmatization for Statistical Constituent Parsing of Italian*

Djamé Seddah[1,2], Joseph Le Roux[3], and Benoît Sagot[2]

[1] Université Paris Sorbonne
[2] INRIA's Alpage project
[3] Université Paris Nord
djame.seddah@paris-sorbonne.fr, leroux@univ-paris13.fr,
benoit.sagot@inria.fr

Abstract. This paper aims at presenting some preliminary results for data driven lemmatization for Italian. Besides intrinsic evaluation for this task, we want to measure its usefulness and adequacy by using our system as input for the task of parsing, following a methodology developed on French. This approach achieves state-of-the-art parsing accuracy without requiring any prior knowledge of the language.

Keywords: Lemmatisation, Statistical Parsing, Italian, PCFGLA, Treebank

1 Introduction

This paper aims at presenting some preliminary results in data driven lemmatization and statistical parsing of Italian based on a methodology we developed on French, a related *cousin* romance language. In our previous work [1, 2], unsupervised word clustering and data driven lemmatization were used as means to alleviate one of Morphologically-Rich Languages' most striking issues [3], namely lexical data sparseness that originates from rich inflections and which is, most of the time, worsen by the small size of syntactically annotated data available for such languages.

Focusing on morphological clustering through lemmatization, and for our first experience in parsing Italian, we decided to use an off-the-shelf lemmatizer based on joint POS tagging and lemmatization model (Morfette, [4]) which was adapted to French with state-of-the-art results regarding the POS and lemmatization tasks [1].

It should be noted that we do not want to perform any post-processing (besides evaluation) to cover cases of deterministic mistakes as we want to evaluate how practical it is for a team of non-native speakers to port a methodology that proved successful on one language to another that shares some of its morphological properties. So far our results are encouraging and show that a large improvement over the baseline can be obtained using a data driven lemmatizer with absolutely no manual interaction.

Maybe more importantly, even with our *naive* tackling of Italian, the lexicon data sparseness reduction induced by our morphological clustering provides state-of-the-art results when used within a PCFG-LA parsing framework.

* Thanks to Alberto Lavelli, Christina Bosco and Fabio Tamburini for answering our questions and making their data available to us. This work is partly funded by the ANR SEQUOIA (ANR-08-EMER-013).

2 Data Driven Lemmatization of Italian

2.1 Training the MORFETTE Model

In order to assign morphological tags and lemmas to words we use the MORFETTE system [4]. It is a sequence labeler that combines the predictions of two classification models (one for morphological tagging and one for lemmatization) at decoding time. While [4] uses MORFETTE's Maximum Entropy models, we use MORFETTE's Averaged Sequence Perceptron models [5] described in [1]. Both classification models incorporate additional features calculated using the MORPH-IT! lexicon [6]. As shown in [7] external lexical data greatly improve Out-of-Vocabulary words handling.

To train MORFETTE, we use the TUT dependency bank made available for the Evalita 2011 Dependency Parsing shared task[4]. MORFETTE being based on a joint POS tagging and lemmatization model, we use both the coarse POS tagset and gold lemma data (resp. columns 3 & 2 of the ITB-DEP) from the whole ITB-DEP as our training set. Besides the trivial replacement of the square and rounded brackets by their usual Penn Treebank counterparts, no specific transformation was applied to the data.

As a morphologically-rich language, Italian exhibits a high level of word form variation. Indeed, we observe from the training set a word/lemma ratio of 1.56. A comparison with the French Treebank (FTB, [8]) shows that both languages displays the same level of word inflection. See Table 1 for a list of these treebanks properties.

Table 1: ITB properties compared to French Treebank

	ITB	FTB
# of tokens	91,720	350,931
# unique word forms	13,092	27,130
# unique lemmas	8362	17,570
ratio words/lemma	1.56	1.54

2.2 Lemmatization Task Results

Baseline As stated earlier, we use the ITB-DEP treebank to extract gold lemmas in addition to gold Part-of-Speech tags. Thus, in order to get an estimate of raw MORFETTE's performance such as in-domain data, we evaluate it on the Evalita dependency parsing gold data. Results are shown in Table 2. Compared to French where lemmatization performances were evaluated on a similar setting, performance range is slightly inferior with a total lemma accuracy of 95.85% on the ITB-DEP test set versus 98.20% on the FTB one. Despite differences in size, this may be explained by a much higher ratio of unknown words in this data set (14.19% vs 4.62%).

[4] `http://www.di.unito.it/~tutreeb/evalita-parsingtask-11.html`

Table 2: POS tagging and lemmatization performance on the ITB-DEP. Evalita 2011

	All	Seen	Unk. (14.19%)
Lemma acc	95.85	97.32	86.95
POS acc	96.47	97.21	92.00
Joint acc	94.47	96.02	85.14

Table 3: Evalita's Dev and Test lemmatization accuracy results

	Evalita	All	Unseen	OOVs (%)
Dev	95.14	83.53	84.77	13.65
Test	94.76	83.78	85.01	18.03

Shared task In order to enforce compatibility with the lemmatization task POS tagset, our original POS tagset was replaced at evaluation time with the one provided by the lemmatization task chair. This choice was made for practical reasons (*i.e.* difficulty to build a mapping between both tagsets without any native Italian speaker in our team) and was possible because POS tags were only used to filter open class words. Evalita's Lemma accuracy is calculated on nouns, adverbs, adjectives and verbs.[5] Table 3 presents our results. The first column contains results provided by the evaluation tool, while the *All* (resp.*Unseen*) column presents lemmatization accuracy calculated for all (resp. absent from training set) tokens.

Our system is ranked #4 where the first three systems outperform our system by at least 3.5 points. One axis of further improvement lies in the way multi-word expressions (MWEs) are handled. In fact, those expressions are marked at the lemma level in the training data while their individual compounds appear as single tokens. This makes the lemmatization task a bit harder for our model. An evaluation without MWEs, simply filtering them out, leads to better results (resp. 95.78 and 95.3 for the Dev and Test set). However, we believe that the highest ratio of OOVs or the Evalita test set, compared to the ITB-DEP test set, has also a significant negative impact on the task. As a matter of fact, we decided to pursue a pure data-driven approach and in this perspective, test set lemmatization performs as if it were out-of-domain. Nevertheless, in-domain evaluation seems to perform as expected by MORFETTE's reported results [4, 1].
In the next section we present preliminary parsing results with data driven lemmatization as a mean of extrinsic evaluation.

3 PCFG-LA Parsing and Lemmatization of Italian

Our main objective is to validate the efficiency of data-driven lemmatization for the task of parsing Italian. For our experiments, we use an in-house parser implementing the CKY algorithm with a coarse-to-fine [9] strategy for context-free grammars with latent annotations (PCFG-LAs) [10]. PCFG-LAs based parsers have been applied to a wide range of languages, among which French [11], German [12] and Italian [13], and always achieve state-of-the-art parsing accuracy.

[5] POS tags: NN, ADV, ADJ and V_*.

Table 4: Predicted Lemma + gold POS results for sentences of length $\leq$40 – – (*p-value (c) & (d), (e) & (c) and (e) & (d)>0.10. All other configurations are statistically significant.*)

Unk. Word Model	LP	LR	F1	Pos Acc	LP	LR	F1	Pos Acc
	No capitalisation				*Original capitalisation*			
Lemma only								
Generic	80.76	81.81	81.28	95.68	80.91	81.95	81.43	95.40
ItalianIG	82.42	83.19	**82.80**	97.29	81.68	82.35	**82.01**	97.25
Lemma + Gold POS								
Generic	84.57	85.22	84.89^{e}	99.96	84.56	85.31	**84.93**^{c}	99.96
ItalianIG	84.59	85.32	**84.95**^{d}	99.96	84.06	84.65	84.35	99.96

Table 5: Predicted Lemma + Predicted POS results for sentences of length $\leq$40 (ITB-DEP coarse grained tagset)

Unk. Word Model	LP	LR	F1	LP	LR	F1
	No capitalisation			*Original capitalisation*		
Predicted POS						
Generic	83.01	83.88	**83.44**	83.37	84.21	**83.79**
ItalianIG	83.05	83.77	83.41	82.96	83.77	83.36

Handling Out-of-Vocabulary (OOV) words in statistical parsing is an under-estimated issue, as parser evaluation is traditionally performed on the Penn Treebank where this problem is almost absent (due to the size of treebank, its homogeneity and because of intrinsic properties of the English language). On the other hand, this issue is of crucial importance for morphologically-rich languages [3], especially when treebanks are small, which is the case here.

Usually, OOV words are assumed to have a POS distribution analogous to rare words in the training set. Hence, in order to reserve a proportion of the probability mass for words unseen in the training set, rare words are replaced with special strings reflecting positional and typographical information (for example suffixes, capitalization, presence/absence of digits) about the replaced word, called signatures. The grammar is learned with these new tokens and during parsing phase, OOV words are replaced with their corresponding signatures.

As none of the authors were familiar with Italian, we decided to use an unsupervised method to detect the useful suffixes for Italian. We use the method introduced in [14] where useful suffixes are extracted from training data and ranked according to their information gain for the task of part-of-speech tagging. Here, suffixes are understood as word endings of length 1, 2 and 3.

3.1 Experimental Protocol

Given the small size of the treebank, our evaluation will be performed with a ten cross-fold validation process using the PARSEVAL metrics: Labelled Precision (LP) and Recall (LR), F-Score (F1) and POS accuracy. All results are given for sentences of length

(strictly) lesser than 41 words. Our PCFG-LA grammars are extracted after 4 refinement iterations.

The architecture we used for French is simple [1, 2]: Each token in the data (training and test) is replaced by a tuple of the form `<lemma,POS>`. Parsers are then trained and tested on the modified data. Original tokens are reinserted before evaluation.

Because the ITB-CONST treebank does not include lemmas, we initially decided to align it with the ITB-DEP treebank at the word level so we could easily generate a lemmatized version of the constituency treebank from gold data. But given the difficulty of aligning both treebanks at the token level (especially when MWEs and traces are annotated differently, the later being erased from our training set), we finally decided to use the MORFETTE model presented in section 2.2 and to reinsert either generated lemmas only or both generated POS tags and lemmas. Results of MORFETTE when evaluated on its own training set are : 97.72% for lemma accuracy, 98.71% for POS accuracy and 97.35% for joint accuracy. Quality can therefore be considered as "good enough" to providing *pseudo* gold lemmas an POS tags. Unfortunately, this entails working with two tagsets: the original reduced one for use in the parser own POS tagging mode (Table 6 and 4) and the ITB-DEP coarse tagset for all predicted `<lemma,POS>` configurations (Table 5) ; consequently, evaluation of predicted POS tags is meaningless by lack of gold ITB-CONST data with the ITB-DEP coarse tagset. All experiment details are available at `http://pauillac.inria.fr/~seddah/Evalita2011`.

Table 6: Baseline PARSEVAL results for sentences of length $\leq$40 – (*p-value (a) & (b)>0.32. All other configurations are statistically significant.*)

Unk. Word Model	LP	LR	F1	Pos Acc
Word only				
ItalianIG	77.30	78.74	78.02	92.82
Generic	79.04	79.96	**79.50**	94.78
Word + Gold Pos				
ItalianIG	83.58	84.29	83.93^{a}	99.98
Generic	83.72	84.22	**83.97**^{b}	99.98

3.2 Results and Discussion

Our baseline results show that the automatic acquisition of signatures for OOV words clearly improves Italian parsing performance in a realistic configuration (F1 of 79.5 vs 78.02). As expected, providing gold POS tags leads to high state-of-the-art results in all configurations[6]. The decrease in OOV words rate achieved by lemmatisation (13.06% in normal mode, 8.48% in lemmatisation without capitalisation and 7.75% otherwise) confirms our hypothesis that morphological clustering (even imperfect) greatly benefits

[6] Lavelli (P.C.) reports 82.88 of F1 on a comparable setting (word+Gold Pos). On the same data, we achieve an F1 of 83.77 using the ItalianIG OOV model (resp. 83.75 with the generic one).

to MRLs parsing, especially in a PCFG-LA framework. Moreover, parsing with predicted lemmas and POS tags drastically improves the global performance compared to our word only baseline. Indeed, this shows that a realistic configuration in this framework performs as well as other reported results with comparable treebanks. In our future work on Italian, we plan to explore more radical forms of word clustering.

References

1. Seddah, D., Chrupała, G., Cetinoglu, O., van Genabith, J., Candito, M.: Lemmatization and statistical lexicalized parsing of morphologically-rich languages. In: Proceedings of SPMRL 2010, Los Angeles, CA (2010)
2. Candito, M., Seddah, D.: Parsing word clusters. In: Proceedings of SPMRL 2010, Association for Computational Linguistics (2010) 76–84
3. Tsarfaty, R., Seddah, D., Goldberg, Y., Kübler, S., Candito, M., Foster, J., Versley, Y., Rehbein, I., Tounsi, L.: Statistical parsing of morphologically rich languages (spmrl): what, how and whither. In: Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages, Association for Computational Linguistics (2010) 1–12
4. Chrupała, G., Dinu, G., van Genabith, J.: Learning morphology with morfette. In: In Proceedings of LREC 2008, Marrakech, Morocco, ELDA/ELRA (2008)
5. Freund, Y., Schapire, R.E.: Large margin classification using the perceptron algorithm. Machine learning **37** (1999) 277–296
6. Zanchetta, E., Baroni, M.: Morph-it!: a free corpus-based morphological resource for the italian language. (2005)
7. Chrupała, G.: Towards a machine-learning architecture for lexical functional grammar parsing. PhD thesis, Dublin City University (2008)
8. Abeillé, A., Clément, L., Toussenel, F. In: Building a Treebank for French. Kluwer, Dordrecht (2003)
9. Charniak, E., Johnson, M.: Coarse-to-fine n-best-parsing and maxent discriminative reranking. In: Proceedings of the 43rd Annual Meeting of the ACL, Barcelona, Spain (2005) 173–180
10. Matsuzaki, T., Miyao, Y., Tsujii, J.: Probabilistic cfg with latent annotations. In: Proc. of ACL-05, Ann Arbor, USA (2005) 75–82
11. Seddah, D., Candito, M., Crabbé, B.: Cross parser evaluation and tagset variation: A French Treebank study. In: Proceedings of the 11th Internation Conference on Parsing Technologies (IWPT'09), Paris, France, Association for Computational Linguistics (2009) 150–161
12. Petrov, S., Klein, D.: Parsing German with latent variable grammars. In: Proceedings of the Workshop on Parsing German at ACL '08, Columbus, Ohio, Association for Computational Linguistics (2008) 33–39
13. Lavelli, A., Corazza, A.: The berkeley parser at the evalita 2009 constituency parsing task. (2009)
14. Attia, M., Foster, J., Hogan, D., Le Roux, J., Tounsi, L., van Genabith, J.: Handling unknown words in statistical latent-variable parsing models for arabic, english and french. In: Proceedings of SPMRL 2010, Association for Computational Linguistics (2010) 67–75